doi: 10.1093/nar/gkz445

MEXPRESS update 2019

Alexander Koch [1,*], Jana Jeschke[2], Wim Van Criekinge[3], Manon van Engeland[1] and Tim De Meyer[3,4]

[1]Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University, 6229 ER Maastricht, the Netherlands, [2]Laboratory of Cancer Epigenetics, Université Libre de Bruxelles, 1070 Brussels, Belgium, [3]Department of Data Analysis and Mathematical Modelling, Ghent University, 9000 Ghent, Belgium and [4]CRIG – Cancer Research Institute Ghent, Ghent University, 9000 Ghent, Belgium

Received February 26, 2019; Revised May 07, 2019; Editorial Decision May 08, 2019; Accepted May 15, 2019

ABSTRACT

The recent growth in the number of publicly available cancer omics databases has been accompanied by the development of various tools that allow researchers to visually explore these data. In 2015, we built MEXPRESS, an online tool for the integration and visualization of gene expression, DNA methylation and clinical data from The Cancer Genome Atlas (TCGA), a large collection of publicly available multi-omics cancer data. MEXPRESS addresses the need for an easy-to-use, interactive application that allows researchers to identify dysregulated genes and their clinical relevance in cancer. Furthermore, while other tools typically do not support integrated visualization of expression and DNA methylation data in combination with the precise genomic location of the methylation, MEXPRESS is unique in how it depicts these diverse data types together. Motivated by the large number of users MEXPRESS has managed to attract over the past 3 years and the recent migration of all TCGA data to a new data portal, we developed a new version of MEXPRESS (https://mexpress.be). It contains the latest TCGA data, additional types of omics and clinical data and extra functionality, allowing users to explore mechanisms of gene dysregulation beyond expression and DNA methylation.

INTRODUCTION

Publicly available cancer multi-omics databases are a valuable asset for scientific research. They help researchers and clinicians decipher the genomic drivers of tumorigenesis, they serve as a valuable resource for the discovery of novel biomarkers and they can be used as independent validation cohorts. One example of such a database is The Cancer Genome Atlas (TCGA). This project, a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute, offers multi-dimensional data on 33 different types of cancer. Analysis of these data has already led to several compelling new insights, such as the discovery of tissue-independent oncogenic molecular signatures and therapeutically actionable alterations (1,2).

User-friendly tools to analyze and visualize large-scale cancer multi-omics data are essential to unlock the full potential of these datasets, as they enable researchers and clinicians without a computational background to explore the data by themselves. The Xena browser (https://xenabrowser.net/), cBioPortal (http://www.cbioportal.org/) (3) and Wanderer (http://maplab.imppc.org/wanderer/) (4) are examples of web-based tools that allow researchers to integrate, analyze and visualize different types of cancer omics data.

In 2015, we created and published MEXPRESS, an online tool for the visualization of TCGA gene expression, DNA methylation and clinical data, as well as the relationships between them (5). One of our goals was to make MEXPRESS as simple to use as possible. Only three steps are required to create a figure: entering a gene, selecting a cancer type and clicking a button. Another major incentive was the lack of proper support for DNA methylation data in existing tools.

DNA methylation consists of the covalent, and therefore reversible, binding of a methyl group to cytosine, one of the four building blocks of DNA. In humans, DNA methylation is almost exclusively restricted to the cytosines of CpG dinucleotides (a cytosine, C, followed by guanine, G). It plays a critical role in the regulation of gene expression, and abnormal DNA methylation patterns are found in virtually every type of human cancer (6). The precise genomic location of DNA methylation is one of the most important factors in the regulatory effect of DNA methylation on gene expression (7), and this is where MEXPRESS distinguishes itself from other available tools.

Using MEXPRESS, researchers can easily investigate the available TCGA DNA methylation data at individual CpGs in relation to their precise genomic location, while at the same time exploring the correlation of the DNA methylation data with gene expression and multiple clinical vari-

[*]To whom correspondence should be addressed. Tel: +31 43 388 42 87; Fax: +31 43 387 46 18; Email: a.koch@maastrichtuniversity.nl

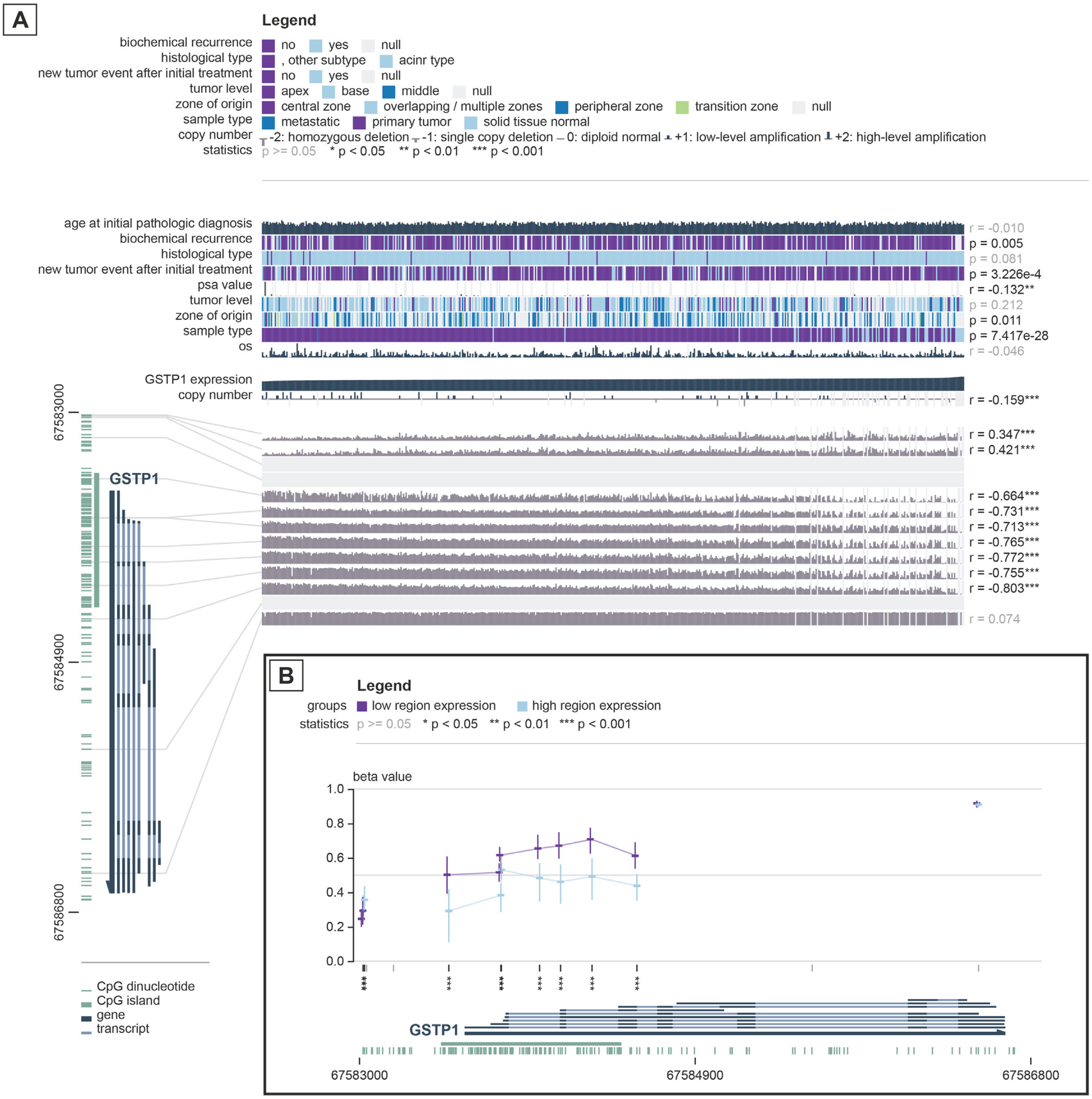

Figure 1. MEXPRESS visualization of the TCGA data for *GSTP1* in prostate adenocarcinoma. (**A**) The default view, in which the samples are sorted by their *GSTP1* expression levels and the samples without expression data were removed. The figure and the statistics on the right hand side show how *GSTP1* expression and promoter DNA methylation are negatively correlated. (**B**) The summary view, a novel feature of MEXPRESS, in which the samples have been divided into two groups based on their *GSTP1* expression level. The horizontal lines at each probe location indicate the median percentage of methylation (β value), whereas the vertical lines mark the range between the 25th and the 75th percentile.

ables. Figure 1A shows the well-known inhibitory effect of promoter hypermethylation on the expression of *GSTP1* in prostate cancer (8). The visual interpretation of the data (less *GSTP1* expression in samples with more promoter methylation) is confirmed by the statistical analysis of the correlation between DNA methylation and gene expression (Pearson correlation coefficients ranging from −0.664 to −0.803 for promoter region probes). We chose to recreate Figure 1 of the original publication (5) to highlight both the similarities (overall layout) and some of the differences (design, ability to zoom in) between the old and the new version of MEXPRESS.

Updates

Several events encouraged us to build a new version of MEXPRESS. First and foremost, MEXPRESS has attracted a sizeable number of users and we still receive positive feedback more than 3 years after the initial publication (5). Between August 2015 and January 2019, a total of 19 623 users visited MEXPRESS and we counted 36 503 sessions for an average of more than 28 sessions per day. Second, the original MEXPRESS back-end pulled data from the TCGA data portal (https://tcga-data.nci.nih.gov/). However, this portal is no longer online and all TCGA data have been migrated to the NCI's Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/), meaning that any updates to the TCGA data were no longer incorporated in the MEXPRESS database.

To address the TCGA data migration, we decided to pull the latest GDC TCGA data from the Xena public data hubs (http://xena.ucsc.edu/public-hubs/). We opted for Xena instead of directly querying GDC because Xena offers much more straightforward data access and pre-merged datasets, meaning that in order to get, for example, all colon adenocarcinoma DNA methylation data only one file needs to be downloaded. The GDC portal on the other hand provides these data as separate files for each sample. Another advantage of using the Xena data hubs is that all data have been processed using the latest human genome assembly (hg38).

While the basic functionality remains the same (enter a gene name and select a cancer type to generate a figure), the new version of MEXPRESS contains multiple incremental improvements, resulting in a substantial update. These updates were inspired by the user feedback we have collected since the initial publication of MEXPRESS, and by a series of in-depth user interviews. Our goal was to improve the existing user interface and user experience and to extend the tool's functionality without sacrificing its user friendliness. Besides a visual overhaul and an update to version hg38 of the human genome, we added the following functionality and data types:

- The ability to filter samples
  Users can filter the samples in the figure by expression, copy number or any of the clinical variables (e.g. select only the samples from male patients where the expression is greater than the median expression value).
- New data types
  We added the following data types to the new version of MEXPRESS: copy number variation, somatic mutations, micro RNA (miRNA) expression, survival data, additional clinical variables and the genomic location of individual CpG dinucleotides. We also included a detailed description of each data type.
- Extra statistical analyses
  We added ANOVA to allow for the comparison of a numerical variable between more than two groups (e.g. to test if there is a difference in expression between the four main tumor stages) and the chi-squared test, to compare a categorical variable between groups (e.g. to test if there is a difference in the gender distribution between normal and tumor samples).
- Summary plot
  Besides the default visualization, users also have the option to create a summary figure, which shows the summarized DNA methylation data for the genomic region selected in the default view (Figure 1B). The samples are split into two or more groups based on the variable by which samples were sorted by (e.g. for a numerical variable, such as expression or patient age, MEXPRESS will create two groups based on the median value of the variable, whereas for a categorical variable, such as sample type, the groups are simply defined by the different categories of this variable). For each Infinium microarray probe visible in the figure, the DNA methylation values are compared between the different sample groups.
- Additional visualization options
  Users can choose which clinical variables they would like to see in the figure and they can zoom in on a specific region of the gene they plotted.
- Download options
  Downloading the figures as SVG or PNG files was already possible in the original version of MEXPRESS, but users can now also download the data that were used to create a figure, as well as the results of the statistical analyses.
- The toolbar
  We collected most functionality and interactivity in a 'toolbar' above the figure. From this toolbar, users can sort and filter samples, select clinical variables, show/hide somatic mutations, switch between the default and summarized view, zoom out, reset the figure, download figures and data, and find detailed information about each data type.
- Step-by-step example
  To guide our users through the new functionality, we added a fully automated step-by-step example analysis to the MEXPRESS homepage.

Copy number variation is one of the data types that have been added to the new version of MEXPRESS. By integrating DNA methylation and copy number data, users can now assess whether a difference in methylation among particular groups represents an actual difference in methylation level or if it merely reflects an underlying difference in copy number. Figure 2 shows a MEXPRESS visualisation of the gene *CDKN2A*, a well-known tumor suppressor (9). Without looking at the copy number data, there appears to be relatively strong positive correlation between *CDKN2A* expression and DNA methylation in the gene body. However, by including the copy number data, we can see how many of the samples with lower levels of *CDKN2A* expression also feature homozygous deletions, which interfere with the estimation of DNA methylation levels. The correlation between DNA methylation and gene expression might therefore be attributed to differences in copy numbers. Users could not have arrived at this insight with the previous version of MEXPRESS, as it lacked the copy number data.

Implementation and data sources

MEXPRESS runs on an Apache server and uses PHP to interact with the back-end MySQL database. Custom python scripts pull the following TCGA data from the Xena public data hubs and upload these data to the MEXPRESS database: gene expression (RNA-seq – HTSeq – FPKM-UQ), DNA methylation (Infinium HumanMethylation450 and HumanMethylation27), gene-level copy number (GIS-

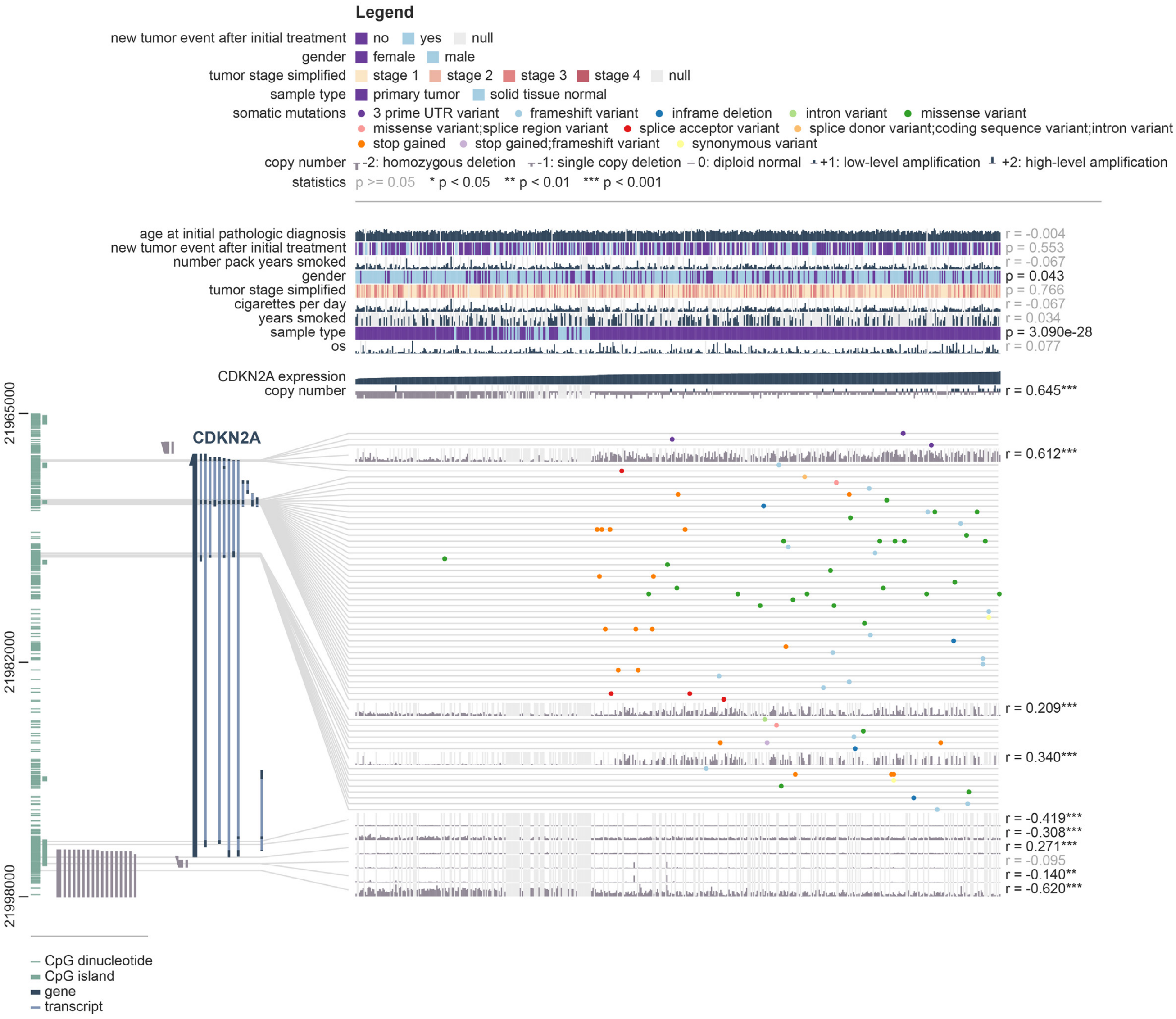

Figure 2. MEXPRESS visualization of the TCGA data for CDKN2A in lung squamous cell carcinoma. This figure shows the complex interplay between gene expression, copy number, DNA methylation and somatic mutations, highlighting the value of the data types that have been added in the new version of MEXPRESS.

TIC2 (10) thresholded), miRNA expression (miRNA-seq RPM), somatic mutations (MuTect2 variant aggregation and masking), phenotype and survival. The copy number data were downloaded from the original TCGA hub (https://tcga.xenahubs.net), while the other data types are downloaded from the GDC hub (https://gdc.xenahubs.net). The reason we used the original TCGA hub for the copy number data is that it provides easy to interpret GISTIC2 thresholded estimated values, representing homozygous deletion (−2), single copy deletion (−1), diploid normal copy (0), low-level copy number amplification (+1) and high-level copy number amplification (+2).

Gene, transcript and exon coordinates were obtained from Ensembl BioMart (https://www.ensembl.org/biomart, Ensembl Genes 94, GRCh38.p12), while miRNA, snoRNA and CpG island coordinates (GRCh38) were obtained from the UCSC genome browser (11). For the Infinium microarray probes, we used the hg38 coordinates from Zhou *et al.* (http://zwdzwd.github.io/InfiniumAnnotation) (12). The breast cancer PAM50 annotation was derived from TCGA's own publication on human breast tumors (13). Finally, the coordinates of the individual CpG dinucleotides are not stored in the MEXPRESS database, but are dynamically requested from the Ensembl REST API (http://rest.ensembl.org/documentation/info/sequence_region).

We performed some minor processing on the TCGA data downloaded from Xena, mostly to ensure consistency in the sample names (all data processing was implemented in R version 3.5.2). We also simplified some of the clinical variables, e.g. we created a new variable 'tumor stage simplified' in which we reduced the many pathologic stages to the four main ones (stages one, two, three and four).

For the MEXPRESS front-end, we used Javascript, jQuery (version 3.2.1), Ajax autocomplete for jQuery (version 1.2.10, https://github.com/devbridge/jQuery-Autocomplete) and the d3.js Javascript library (version 5.5.0, http://d3js.org/). When a user downloads a figure, the SVG image is converted into a PNG image using Inkscape, an open source vector graphics editor (http://www.inksca-pe.org/).

All MEXPRESS code (back-end, front-end and data processing) can be cloned or downloaded from this GitHub repository: https://github.com/akoch8/mexpress.

CONCLUSION

The number of users proves that MEXPRESS, through its ease of use and unique, integrative data overview, found its place in the toolbox of many researchers. By combining a comprehensive visualization and statistical analysis in a single figure, MEXPRESS helps researchers quickly identify dysregulations and their clinical relevance in cancer. With this major, feedback-driven update, we aim to consolidate MEXPRESS's place in the set of open source web tools available to researchers and clinicians.

ACKNOWLEDGEMENTS

This work was supported by a SU2C-DCS International Translational Cancer Research Dream Team Grant (Stand Up To Cancer (SU2C)-AACR-DT1415, MEDOCC).

FUNDING

SU2C-DCS International Translational Cancer Research Dream Team Grant [Stand Up To Cancer (SU2C)-AACR-DT1415, MEDOCC]. Funding for open access charge: SU2C-DCS International Translational Cancer Research Dream Team Grant [Stand Up To Cancer (SU2C)-AACR-DT1415, MEDOCC].

Conflict of interest statement. None declared.

REFERENCES

1. Ciriello,G., Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.
2. Sanchez-Vega,F., Mina,M., Armenia,J., Chatila,W.K., Luna,A., La,K.C., Dimitriadoy,S., Liu,D.L., Kantheti,H.S., Saghafinia,S. *et al.* (2018) Oncogenic signaling pathways in the cancer genome atlas. *Cell*, **173**, 321–337.
3. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, l1.
4. Diez-Villanueva,A., Mallona,I. and Peinado,M.A. (2015) Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics Chromatin*, **8**, 22.
5. Koch,A., De Meyer,T., Jeschke,J. and Van Criekinge,W. (2015) MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics*, **16**, 636.
6. Herman,J.G. and Baylin,S.B. (2003) Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med.*, **349**, 2042–2054.
7. Koch,A., Joosten,S.C., Feng,Z., de Ruijter,T.C., Draht,M.X., Melotte,V., Smits,K.M., Veeck,J., Herman,J.G., Van Neste,L. *et al.* (2018) Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.*, **15**, 459–466.
8. Millar,D.S., Ow,K.K., Paul,C.L., Russell,P.J., Molloy,P.L. and Clark,S.J. (1999) Detailed methylation analysis of the glutathione S-transferase pi (GSTP1) gene in prostate cancer. *Oncogene*, **18**, 1313–1324.
9. Liggett,W.H. Jr and Sidransky,D. (1998) Role of the p16 tumor suppressor gene in cancer. *J. Clin. Oncol.*, **16**, 1197–1206.
10. Mermel,C.H., Schumacher,S.E., Hill,B., Meyerson,M.L., Beroukhim,R. and Getz,G. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.
11. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
12. Zhou,W., Laird,P.W. and Shen,H. (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.*, **45**, e22.
13. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.